INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 3, Issue 6)
Available online at www.ijariit.com

Big Data Analysis Model for Transformation and Effectiveness in the e-Governance Service from Relational Database

Ashish Dhingra
Centre for Development and Advanced Computing, Mohali, Punjab
ashish.dhingra@nic.in

Iqbal Singh
Centre for Development and Advanced Computing, Mohali, Punjab
iqbalme@gmail.com

Abstract: Relational Database Management Systems are frequently being used in various e-Governance projects in India, but RDBMS are not suitable for OLAP (Online analytical Processing). By using RDBMS, the large number of the volume have been collected and processed in e-Governance projects. Big Data framework based processing has been implemented to study various parameters to provide faster and better communication, retrieval of data and utilization of information to its users in e-Governance projects. This research focuses on how we can use Big Data framework for effective implementation of e-Governance Projects.

Keywords: Big Data, Hadoop, e-Governance, e-District, Migration.

I. INTRODUCTION

In last five years of so, Government of India is pushing towards e-Governance, which has led to the transformation of various offline applications to online applications. Most of the applications are capturing real time data. Even if the mode of submission is offline now Management Information System (MIS) is need of the hour. So various State Governments, Departments are now adopting to e-Governance framework. In nutshell, this all has resulted in the generation of huge data. In Today's world, relevant date i.e. information is a superpower. Therefore we must have some mechanism to use that data for taking various policy decisions of the Government. Analytics plays an important role in strategic planning and implementation of various schemes of Government. Most of the e-Governance projects in India use traditional Relational Database Management System (RDBMS), which fulfills their requirement of storing and fetching data for various reports etc. These RDBMS have their own limitations, thus need of Big Data Analytics was envisaged very well.

II. BACKGROUND & DEFINITION

First, the data size has increased tremendously in the range of petabytes—one petabyte = 1,024 terabytes. RDBMS finds it challenging to handle such huge data volumes. To address this, RDBMS added more central processing units (or CPUs) or more memory to the database management system to scale up vertically. Second, the majority of the data comes in a semi-structured or unstructured format from social media, audio, video, texts, and emails. However, the second problem related to unstructured data is outside the purview of RDBMS because relational databases just can't categorize unstructured data. They're designed and structured to accommodate structured data such as weblog sensor and financial data. Also, "big data" is generated at a very high velocity. RDBMS lacks in high velocity because it's designed for steady data retention rather than rapid growth. Even if RDBMS is used to handle and store "big data," it will turn out to be very expensive. As a result, the inability of relational databases to handle "big data" led to the emergence of new technologies.

e-Governance is one of the forms of Good Governance. e-Governance in India has taken a revolutionary change after Government has introduced National e-Governance Plan. This approach has the potential of enabling huge savings in costs through sharing of core and support infrastructure, enabling interoperability through standards, and of presenting a seamless view of Government to citizens.

The National e-Governance Plan (NeGP) takes a holistic view of e-Governance initiatives across the country, integrating them into a collective vision, a shared cause. Around this idea, a massive countrywide infrastructure reaching down to the remotest of

villages is evolving, and large-scale digitization of records is taking place to enable easy, reliable access over the internet. The ultimate objective is to bring public services closer home to citizens, as articulated in the Vision Statement of NeGP.

"Make all Government services accessible to the common man in his locality, through common service delivery outlets, and ensure efficiency, transparency, and reliability of such services at affordable costs to realize the basic needs of the common man".

The Government approved the National e-Governance Plan (NeGP), comprising of 27 Mission Mode Projects and 8 components, on May 18, 2006. In the year 2011, 4 projects - Health, Education, PDS, and Posts were introduced to make the list of 27 MMPs to 31Mission Mode Projects (MMPs). The Government has accorded approval to the vision, approach, strategy, key components, implementation methodology, and management structure for NeGP. However, the approval of NeGP does not constitute financial approval(s) for all the Mission Mode Projects (MMPs) and components under it. The existing or ongoing projects in the MMP category, being implemented by various Central Ministries, States, and State Departments would be suitably augmented and enhanced to align with the objectives of NeGP.
In order to promote e-Governance in a holistic manner, various policy initiatives and projects have been undertaken to develop core and support infrastructure. The major core infrastructure components are:

1. State Data Centres (SDCs)
2. State Wide Area Networks (S.W.A.N)
3. Common Services Centres (CSCs) and middleware gateways i.e National e-Governance Service Delivery Gateway (NSDG)
4. State e-Governance Service Delivery Gateway (SSDG), and Mobile e-Governance Service Delivery Gateway (MSDG).

The important support components include Core policies and guidelines on Security, HR, Citizen Engagement, Social Media as well as Standards related to Metadata, Interoperability, Enterprise Architecture, Information Security etc. New initiatives include a framework for authentication, viz. e-Pramaan and G-I cloud, an initiative which will ensure benefits of cloud computing for e-Governance projects.

Big Data takes care of Large and complex data which is difficult to process using traditional applications. Extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions. It concentrates on 4Vs of data, which are given below:

- Volume (scale of data)
- Variety (different forms of data)
- Velocity (analysis of streaming data)
- Veracity (uncertainty of data)

The Big Data can be utilized by the Government for Sentiment Analysis, Machine Learning and enriching the Data Mining capabilities of existing Data Ware housing and Business Analytics systems. Traditional RDBMS cannot handle big data sets so document Databases known as NoSQL databases are used to handle them.

III. MIGRATING e -GOVERNANCE PROJECT IN BIG DATA FRAMEWORK

This research focuses on how we can migrate existing e-Governance relational database into Big data framework. Following tasks were performed:

1. Install Hadoop on single node cluster. Operating System used during research is Windows 10. See Apache Hadoop Wiki Page for step by step installation. Once everything is set, you can run start-dfs.cmd & start-yarn.cmd command from command prompt to run all Hadoop services. It will start following:-

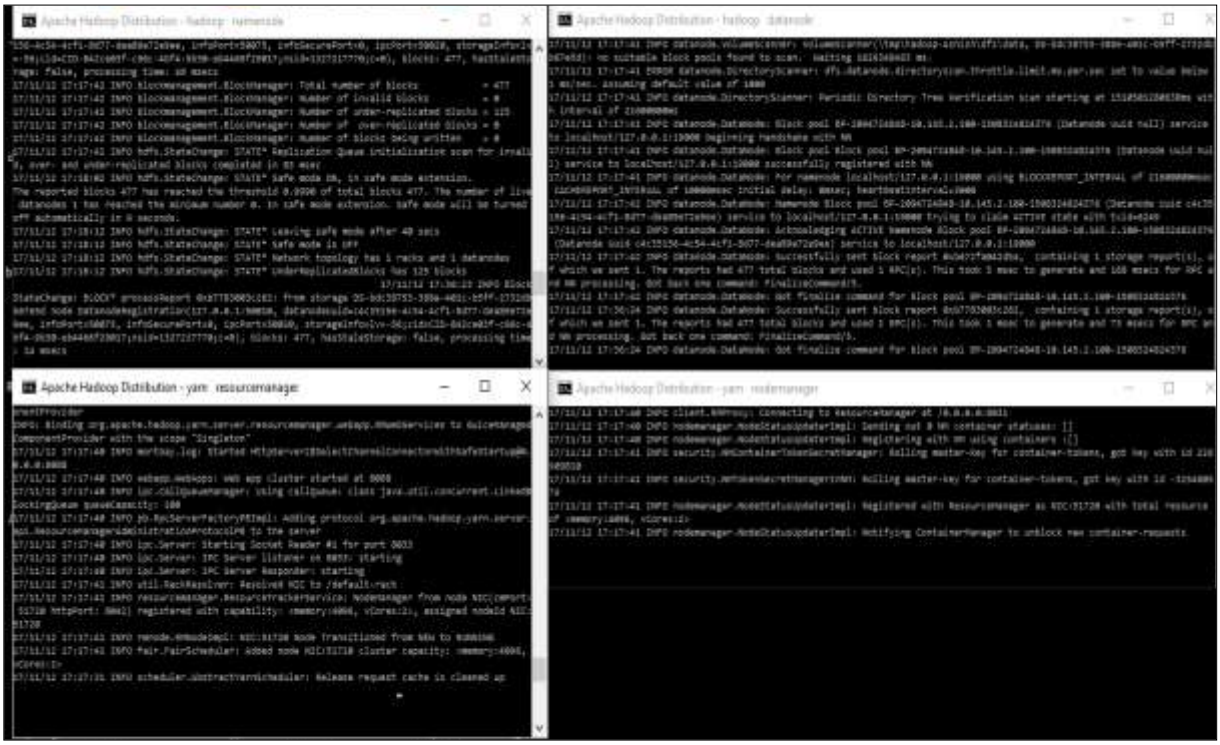
    a. **NameNode;** The NameNode is the centerpiece of an HDFS file system. It keeps the directory tree of all files in the file system, and tracks where across the cluster the file data is kept. It does not store the data of these files itself. Client applications talk to the NameNode whenever they wish to locate a file, or when they want to add/copy/move/delete a file. The NameNode responds the successful requests by returning a list of relevant DataNode servers where the data lives. The NameNode is a Single Point of Failure for the HDFS Cluster. HDFS is not currently a High Availability system. When the NameNode goes down, the file system goes offline. There is an optional Secondary NameNode that can be hosted on a separate machine. It only creates checkpoints of the namespace by merging the edits file into the fsimage file and does not provide any real redundancy. Hadoop 0.21+ has a Backup NameNode that is part of a plan to have an HA name service, but it needs active contributions from the people who want it (i.e. you) to make it Highly Available.

    b. **DataNode:** A DataNode stores data in the [Hadoop FileSystem]. A functional filesystem has more than one DataNode, with data replicated across them. On startup, a DataNode connects to the NameNode; spinning until that service comes up. It then responds to requests from the NameNode for filesystem operations. Client

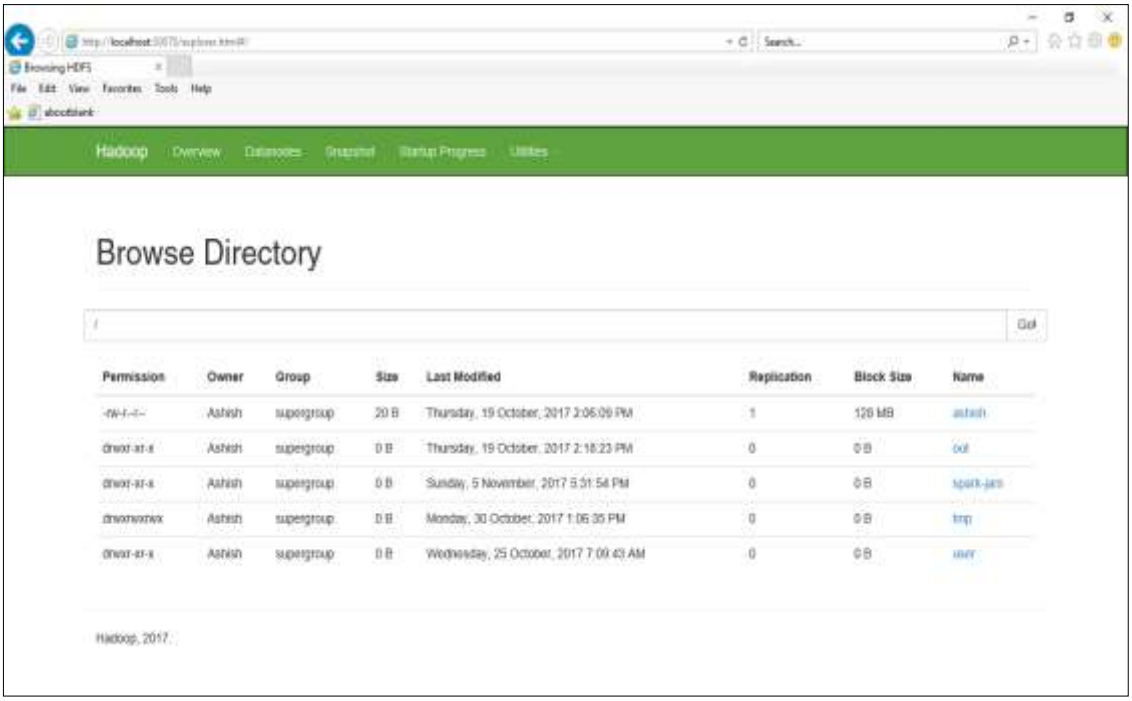
applications can talk directly to a DataNode, once the NameNode has provided the location of the data. Similarly, MapReduce operations farmed out to TaskTracker instances near a DataNode, talk directly to the DataNode to access the files. TaskTracker instances can, indeed should, be deployed on the same servers that host DataNode instances, so that MapReduce operations are performed close to the data. DataNode instances can talk to each other, which is what they do when they are replicating data.

c. **Resource Manager:** The **Resource Manager** (one per cluster) is the master. It knows where the slaves are located (Rack Awareness) and how many resources they have. It runs several services, the most important is the **Resource Scheduler** which decides how to assign the resources.

d. The **Node Manager** (many per cluster) is the slave of the infrastructure. When it starts, it announces himself to the Resource Manager. Periodically, it sends a heartbeat to the Resource Manager. Each Node Manager offers some resources to the cluster. Its resource capacity is the amount of memory and the number of vcores. At run-time, the Resource Scheduler will decide how to use this capacity: a **Container** is a fraction of the NM capacity and it is used by the client for running a program.



1. Hadoop file system is the heart of Apache Hadoop project. The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. HDFS was originally built as infrastructure for the Apache Nutch web search engine project.
   We will create a directory in hdfs and migrate data using Sqoop tool to import data from relational database to hdfs. Contents stored in HDFS can be browsed through a browser at http://localhost:50070/explorer.html#/. Localhost can be replaced by Server IP address, which hosts Hadoop Installation.



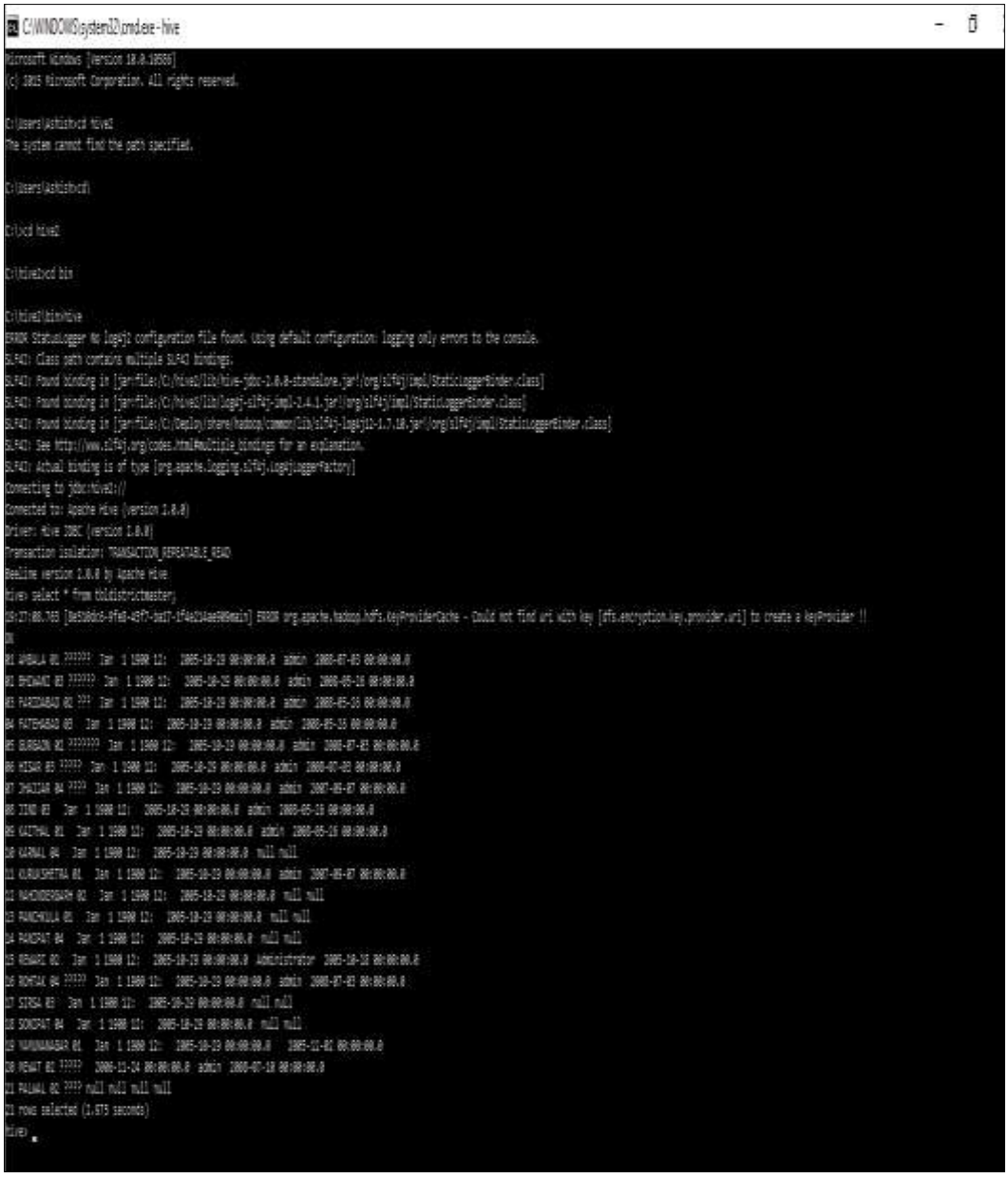
2. Install Sqoop to import e-District data, which is e-Governance project to Hadoop file system. It is the beauty of Sqoop that you can directly connect to your existing RDBMS using Java Database Connectivity (JDBC).

```
sqoop list-databases --connect jdbc:sqlserver://localhost:1433 --username hadoop --password hadoop
```

In the e-District project, Microsoft Sql Server is being used as Relational Database Management System. Microsoft Sql Server runs on 1433 port. For other RDBMS, you need to locate for the relative port. Once you are connected to the database, we can give import command to import the database into HDFS.

```
sqoop import --connect "jdbc: sqlserver://localshot:1433;database=edistrict;username=hadoop;password=hadoop" --table tbldistrictmaster --hive-import
```

The Apache Hive ™ data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL. The structure can be projected onto data already in storage. A command line tool and JDBC driver are provided to connect users to Hive. With the above command, table tbldistrictmaster will get imported in Hive and we will Hive to query this table in Hadoop framework. Similar all tables will be imported in Hive of the e-District database.



IV. RESULTS & FINDINGS

Based on the research, following positive outcomes were found:

1. **Open Source:** We have seen in this research that all tools used are Open Source, therefore you need not pay anything for licensing, where as in traditional RDBMS, licensing cost is very high.
2. **Compressed Data Storage:** Hive can store data itself in a compressed format, with ORC format support; we can store even a specific column as compressed one.
3. **Distributed Computing:** The UCP of Hadoop is Distributed Computing. Data can be distributed various data nodes (Slaves). A name node (Master) can send a request to all data nodes and compile the result. This technique is also known as Map-Reduce. Larger the no. of nodes, faster data will be processed. By this technique, huge data of e-District databases can be processed easily as compared to RDBMS. In our research, we have configured Hadoop on Single Node. On Multinode cluster, performance will be much better.
4. **Data Type Support:** Traditional RDBMS can handle only structured data, whereas Hadoop can handle unstructured data as well.

V. CONCLUSION

Traditional databases are well suited for Online Transaction Processing (OLTP), but when it comes to fetching huge data and doing analytics on it (Online/Offline Analytical Processing), Relational Databases has its own limitations. Therefore Big Data framework provides flexibility in fetching huge amount of data and analyzes data. The best part is that it is Open Source, so doesn't involve any extra cost and this research shows that you can port your existing data in Hadoop framework. If used in e-Governance projects, Big Data Analytics can help in strategic planning and policy level decisions.

REFERENCES

[1] Preet Navdeep, Manish Arora and Neeraj Sharma (2016). Role of Big Data Analytics in Analyzing e-Governance Projects. GIAN JYOTI E-JOURNAL, Volume 6, Issue 2 (Apr-Jun 2016).

[2] Dipika Thakare (2016). A Survey of Migration from Traditional Relational Databases towards To New Big Data TechnologyInternational Journal of Innovative and Emerging Research in Engineering Volume 3, Issue 2, 2016.

[3] Nishant Agnihotri and Aman Kumar Sharmat(2015). International Journal of Innovations & Advancement in Computer Science IJIACS ISSN 2347 – 8616 Volume 4, Special Issue March 2015.

[4] Karthik Kambatla, Giorgos Kollias, Vipin Kumar, Ananth Grama (2014). Trends in Big Data Analytics Journal of Paralleland Distributed Computing.

[5] Anthony G. Picciano. The Evolution of Big Data and Learning Analytics in American Higher Education; Article in Journal of Asynchronous Learning Network · June 2012.

[6] https://wiki.apache.org/hadoop/NameNode accessed Novemeber 2017

[7] https://wiki.apache.org/hadoop/Hadoop2OnWindows accessed Novemeber 2017

[8] http://meity.gov.in/divisions/national-e-Governance-plan accesed November 2017

